

PATENT

A57112/02

SUND-002/02

EEPROM WITH SPLIT GATE SOURCE SIDE INJECTION

INTRODUCTIONRelated Applications

This application is a continuation-in-part of U.S. Serial Number 08/193,707 filed February 9, 1994, which in turn is a divisional of U.S. Serial Number 07/820,364, filed January 14, 1992, now U.S. Patent 5,313,421 issued May 17, 1994.

Technical Field

This invention pertains to semiconductor memory cells and arrays, more particularly to electrically erasable programmable read only memories.

Background

Erasable programmable read only memories (EPROMs) and electrically erasable programmable read only (EEPROMs) are well known in the art. These devices have the ability to store data in non-volatile fashion, while also being capable of being erased and rewritten as desired. EPROM devices are typically erased by exposing the integrated circuit device to ultraviolet radiation, while EEPROMs allow erasure to be performed electrically.

One form of EEPROM device includes a so-called "split-gate" electrode, in which the control gate includes a first portion overlaying a floating gate and a second portion directly overlaying the channel. Such a split gate structure is described in a 5-Volt-Only Fast-Programmable Flash EEPROM Cell with a Double Polysilicon Split-Gate Structure by J. Van Houdt et al, Eleventh IEEE Non-Volatile Semiconductor Workshop,

1 February 1991, in which charge is injected into the floating
2 gate from the source side of the cell. U.S. Patent 4,652,897
3 describes an EEPROM device which does not utilize a split-gate,
4 but which also provides injection to the floating gate from the
5 source side of the device.

6 As described in the above referenced U.S. Patent 4,652,897,
7 memory cells are typically arranged in an array, as is well
8 known in the art. One form of such an array utilizes buried
9 diffusions, in which source and array regions are covered with
10 a fairly thick layer of insulating material. This is shown for
11 example, in U.S. Patents 4,151,020; 4,151,021; 4,184,207; and
12 4,271,421. Such buried diffusion devices often utilize a
13 virtual ground approach, in which columns connecting the sources
14 of a first column of memory cells also serves to connect drains
15 of an adjacent column of memory cells.

16 While many EEPROM devices utilize two layers of
17 polycrystalline silicon, one for the formation of the floating
18 gate, and the other for the formation of the control gate and
19 possibly electrical interconnects, other EEPROM devices utilize
20 three layers of polycrystalline silicon. For example, U.S.
21 Patent 4,302,766 provides a first polycrystalline silicon layer
22 for the floating gate, a second polycrystalline silicon layer
23 for the control gate, and a third polycrystalline silicon layer
24 coupled through an erase window to a portion of the first
25 polycrystalline silicon layer for use during erasure of the
26 cell. U.S. Patent 4,331,968 also uses a third layer of
27 polycrystalline silicon to form an erase gate, while U.S. Patent
28 4,462,090 forms an addressing gate electrode utilizing a third
29 layer of polycrystalline silicon. U.S. Patent 4,561,004 and
30 4,803,529 also use three layers of polycrystalline silicon in
31 their own specific configurations.

32 Japanese Patent Publication 61-181168 appears to utilize
33 three layers of polycrystalline silicon to provide additional
34 capacitive coupling to the floating gate. Japanese Patent

1 Publication 63-265391 appears to pertain to a buried diffusion
2 array, possibly utilizing virtual grounds.

3 European Patent Application 0373830 describes an EEPROM in
4 which two polycrystalline silicon layers are used, with the
5 second layer of polycrystalline silicon having two pieces, one
6 of which provides the erase function, and one of which provides
7 the steering function.

8 "A New Flash-Erase EEPROM Cell With a Sidewall Select-Gate
9 on its Source Side" by K. Naruke et al. IEDM-89-603 and U.S.
10 Patent 4,794,565 describe an EEPROM utilizing a side wall select
11 gate located on the source side of the field effect transistor.

12 "EPROM Cell With High Gate Injection Efficiency" by
13 M. Kamiya et al. IEDM 82-741, and U.S. Patent 4,622,656 describe
14 an EEPROM device in which a reduced programming voltage is
15 provided by having a highly doped channel region under the
16 select gate, and the channel region under the floating gate
17 being either lightly doped or doped to the opposite conductivity
18 type, thereby providing a significant surface potential gap at
19 the transition location of the channel.

20 In recent years there has been significant interest in
21 producing high capacity FLASH memory devices which use split-
22 gate, source-side hot electron programming, in place of the more
23 conventional drain-side channel hot electron (CHE) mechanism.

24 The reasons for this include its inherently lower write
25 power requirement (1/10th that of CHE or less), facilitating low
26 voltage operation and higher write speeds via increased
27 parallelism. In addition, the split gate structure is not
28 susceptible to "overerase" related problems (a problem for
29 single gate FLASH memories such as ETOX), and does not
30 experience programming difficulty due to strong overerase, which
31 can hinder programming after an erasure operation in split-gate
32 CHE programming devices.

33 In view of these benefits, SunDisk Corporation has patented
34 FLASH memory cell and array variants which use source side

1 injection integrated with SunDisk's proprietary thick oxide,
2 poly-to-poly erase tunneling technology, to make a highly
3 scalable, reliable, low power programming cell (D.C. Guterman,
4 G. Samachiasa, Y. Fong and E. Harari, U.S. Patent No.
5 5,313,421).

6 The concept of a multi-bit storage non-volatile cell using
7 a split gate structure was described by G.S. Alberts and H.N.
8 Kotecha (Multi-bit storage FET EAROM cell, IBM Technical
9 Disclosure Bulletin, Vol. 24 No. 7A, p. 3311, Dec. 1981). They
10 describe a two-poly, three transistor element-in-series cell, in
11 which the center transistor's channel is controlled directly by
12 the poly2 control gate (which also serves as the cell select
13 gate), and each of the two end transistor channels are
14 controlled by corresponding poly1 floating gates, which in turn
15 are capacitively coupled to the control gate, thereby realizing
16 a plurality of bits in the one physical cell structure.

17 Recently, at the 1994 IEDM, Bright Microelectronics along
18 with Hyundai presented a similar dual-bit split-gate cell,
19 integrated into a contactless, virtual ground array, and using
20 source side injection programming (Y.Y. Ma and K. Chang, U.S.
21 Patent No. 5,278,439 - referred to henceforth as the Ma
22 approach). One structural difference here from the IBM approach
23 is their separation of the capacitively coupling control gates,
24 which are formed in poly2, and the select gate, which is formed
25 in poly3.

26 In the Ma approach, they use "conventional" negative
27 control gate driven tunneling through an ultra-thin poly1 gate
28 oxide (about 100Å or less). This erase approach poses some
29 serious limitations. Erase of one of the two storage
30 transistors uses floating gate to drain tunneling through the
31 ultra-thin oxide, accomplished by biasing the drain to 7v and
32 corresponding control gate to -10v. Because both of these lines
33 run perpendicular to the select gate, this forces a block of
34 cells which are to be simultaneously erased (e.g. a sector) to

1 be bit line oriented, as opposed to the more conventional word
2 line (select gate) oriented block; i.e. its sector must be
3 column organized and thus it cannot be row organized. (For
4 example, a sector could be two columns of floating gates
5 straddling a bit line/diffusion, including the right hand
6 floating gates of the left side cells' floating gate pair plus
7 the left hand floating gates of the right side cells.) This
8 leads to the following disadvantages in the Ma implementation:
9 (1) Limited to column sector architecture; i.e. cannot readily
10 support the higher read performance row oriented sector
11 architecture. (Since here, within a sector, both erase anode
12 and corresponding control gates run perpendicular to row line
13 direction, this precludes the massively parallel "chunk"
14 implementation of the row oriented sector, which can
15 simultaneously access large numbers of cells within that
16 sector).
17 (2) Requires ultra-thin, approximately 100Å, tunneling oxide,
18 imposing following limitations:
19 * Scaling limitation associated with pushing the limits
20 of usable oxide thicknesses, plus the additional area needs
21 associated with maintaining adequate coupling requirements,
22 which must combat the inherently high capacitance per unit area
23 of such a thin oxide;
24 * A myriad of potential retention/reliability problems
25 inherent to using ultra-thin oxide, combined with the parasitic
26 band-to-band tunneling/hole injection associated with the high
27 substrate fields adjacent to the diffusion anode; and
28 * Negative gate bias requirements on control gate, to
29 limit band-to-band injection problems, impose process and
30 circuit complexity, plus potentially more layout area
31 requirement.
32
33

SUMMARY OF THE INVENTION

1 In accordance with the teachings of this invention, novel
2 memory cells are described utilizing source-side injection.
3 Source-side injection allows programming utilizing very small
4 programming currents. If desired, in accordance with the
5 teachings of this invention, to-be-programmed cells along a
6 column are programmed simultaneously which, due to the small
7 programming current required for each cell, does not require an
8 unacceptably large programming current for any given programming
9 operation. In one embodiment of this invention, the memory
10 arrays are organized in sectors with each sector being formed of
11 a single column or a group of columns having their control gates
12 connected in common. In one embodiment, a high speed shift
13 register is used in place of a row decoder in order to serially
14 shift in the data for the word lines, with all of the data for
15 each word line of a sector being contained in the shift register
16 on completion of its serial loading. In one embodiment,
17 additional speed is achieved by utilizing a parallel loaded
18 buffer register which receives data in parallel from the high
19 speed shift register and holds that data during the write
20 operation, allowing the shift register to receive serial loaded
21 data during the write operation for use in a subsequent write
22 operation. In one embodiment, a verification is performed in
23 parallel on all to-be-programmed cells in a column and the bit
24 line current monitored. If all of the to-be-programmed cells
25 have been properly programmed, the bit line current will be
26 substantially zero. If bit line current is detected, another
27 write operation is performed on all cells of the sector, and
28 another verify operation is performed. This write/verify
29 procedure is repeated until verification is successful, as
30 detected by substantially zero bit line current.

31 Among the objectives of the novel cells constructed in
32 accordance with this invention are avoidance of programming
33 limitations such as:

34

1. High Channel Currents (Power) required for Programming.
2. High Drain Voltage Requirements, which increase with increased levels of erasure.
3. Loss of Read Performance associated with an increase in Programming Efficiency via Heavy Channel doping.
4. Program Wearout Associated with Maintaining a High Drain Bias on Cells exposed to this bias, including both those cells targeted for programming and those cells not targeted but still exposed to the voltage.

In an alternative embodiment of this invention, a multi-bit memory cell is taught utilizing a 3-poly, 3 transistor element-in-series cell in which the center transistor's channel is controlled directly by the poly 3 control gate (which serves as both a cell select gate and erase anode) and each of the two end transistor channels are controlled by corresponding poly1 floating gates, which in turn are capacitively coupled to the poly 2 control or steering gates, thereby realizing a plurality of bits in the one physical cell structure.

The multi-bit cell contains two bits per unit memory cell, coming from two floating gate portions, each having their own control gate (which, in the virtual ground array, runs parallel to the bits lines), and sharing one select gate, placed physically between them (which, in the virtual ground array, runs perpendicular to the bit lines). The diffusion BN+ source/drains straddle the two floating gates, on their opposite facing channel edges to those adjacent the select gate/transfer channel.

Unlike a single floating gate cell, because here the two floating channels lie in a series configuration, the programmed threshold voltage level of each floating gate must be limited in its upper value in order to be readable (similarly to the Toshiba NAND cell). In this way, either floating gate channel

can be unconditionally turned on (i.e. independent of its stored state) when appropriate bias is applied to its corresponding control gate, when reading the state of the other floating gate.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1a, 1b, and 1c, are cell layout, cross-sectional diagram, and equivalent circuit schematic of one embodiment of this invention;

Figure 1d is a plan view of one embodiment of an array consisting of a plurality of cells of Figures 1a-1c;

Figure 1e is a block diagram depicting a memory array organized by sectors, with appropriate control circuitry;

Figure 1f depicts the operation of one embodiment of a memory array organized by sectors as shown in Figure 1e;

Figure 1g is a plan view depicting an alternative array embodiment utilizing cells depicted in Figures 1a-1c;

Figure 2a is a cross-sectional view depicting an alternative embodiment of this invention similar that of Figure 1b;

Figure 2b is a plan view of one embodiment of an array of memory cells constructed utilizing cells depicted in the cross-sectional view of Figure 2a;

Figure 2c is a diagram depicting the organization and operating condition of an array such as that of Figure 2b;

Figure 3 is a graph depicting the operation of a memory cell of Figure 1b;

Figure 4 depicts the electrical field distribution along channels of the device of Figure 5;

Figure 5 is a cross-sectional view one embodiment of a 2-poly cell of this invention;

Figure 6 is a cross-sectional view of another embodiment of a 2-poly cell of this invention;

Figure 7a is a plan view depicting a portion of a process sequence utilized in accordance with one embodiment of this invention;

Figure 7b is a cross-sectional view of the embodiment shown in the plan view of Figure 7a;

Figure 8 is a cross-sectional view depicting a fabrication step suitable for use in accordance with the teachings of this invention;

Figures 9a and 9b are top and cross-sectional views, respectively of one embodiment of a multiple-bit memory cell structure of this invention;

Figure 10a is a schematic diagram of one multi-bit cell of this invention;

Figure 10b is a circuit diagram depicting one embodiment of an array of multiple-bit memory cells of this invention, such as those of Figures 9A and 9B;

Figure 10c is a circuit diagram depicting one embodiment of an array of cells as shown in Figure 10b plus segment decode transistor matrix for both bit lines and steering lines;

Figures 11a through 11e are detailed top and cross-sectional views; and

Figures 12a-12f are cross sectional views depicting fabrication steps suitable for use in fabricating multi-bit memory cells in accordance with this invention.

DESCRIPTION OF SPECIFIC EMBODIMENTS

The cell layout, cross-sectional diagram and equivalent circuit schematic of one embodiment are shown in Figures 1a, 1b, and 1c, respectively. Similar reference numerals have been used in Figures 1a, 1b, and 1c. Referring to the cross-sectional view of Figure 1b, this embodiment of the novel EEPROM cell, 101, of this invention includes a buried source region 102 and a buried drain region 103, each being buried by a relatively thick layer of dielectric 104 and 105, respectively. Channel

1 region 106 is divided into two portions, a first portion 106-1
2 which is influenced by the third layer polycrystalline silicon
3 109 and which forms a select gate, and a second portion 106-2
4 which is influenced by floating gate 107 formed of a first layer
5 of polycrystalline silicon and which, in turn, is influenced by
6 control gate 108 formed of a second layer polycrystalline
7 silicon. As is well known in the art, suitable dielectric
8 layers such as thermally grown oxide are located between channel
9 106 and polycrystalline silicon layer 109 and polycrystalline
10 silicon layer 107. Similarly, suitable dielectric layers such
11 as oxide or composite oxide/nitride are formed between the three
12 layers of polycrystalline silicon. Polycrystalline metal
13 silicide can be used in place of one or more of the
14 polycrystalline silicon layers 108 and 109. If desired, a
15 highly-doped P+ region 120 is used within channel 106-2 adjacent
16 buried drain region 103. This region 120 is formed, for
17 example, as a double diffused MOS (DMOS) region in order to
18 establish the threshold voltage V_t of the memory transistor
19 including channel 106-2. This helps to provide a stable
20 threshold voltage, even though the amount of charges trapped in
21 the gate oxide layer in the vicinity of the gap between 106-1
22 and 106-2 tends to increase with a large number of programming
23 cycles.

24 An example of operating conditions and levels associated
25 with the embodiment of Fig. 1b are shown in Table 1. High
26 efficiency programming comes about by the ability to
27 simultaneously create a high field region in channel 106-2 under
28 the floating gate, which under the bias conditions of Table 1
29 occur near the gap between channels 106-1 and 106-2 (see above
30 mentioned IEDM article of Kamiya for theory) while maintaining
31 a low channel/current. Since this high field causes electron
32 injection to floating gate 107 near the source side of channel
33 106-2, this type of operation is termed "source-side" injection.
34 This mechanism provides high efficiency, low power programming

1 by maintaining a low channel current via word line 109
2 throttling by using a bias operating near channel threshold,
3 VT_{p3} . A major attribute of this type of operation is that it
4 allows for a high drive condition in floating gate channel 106-2
5 under the floating gate (in fact it thrives on it), offering
6 high-performance read, without degrading programming
7 performance. This is because the very weak drive condition on
8 the select transistor of channel 106-1 is established via the
9 throttling mentioned above to achieve the high fields in the
10 vicinity of the poly 3/poly 1 gap. These fields accelerate the
11 electrons to sufficiently energetic levels (i.e. > 3.1 eV) to
12 surmount the Si/SiO₂ interface barrier at the source side of
13 floating gate 107. Furthermore, there is a significant vertical
14 component to that field (i.e. normal to the Si/SiO₂ surface)
15 driving the electrons up to the surface of channel 106, and
16 thereby assisting the injection into floating gate 107. No read
17 performance penalty is incurred to establish this high field
18 condition. This is in stark contrast to conventional drain side
19 programming, wherein efficient program requires strong channel
20 saturation which shuns high floating gate channel drives, strong
21 overerase, or a weakly turned on series select transistor.
22 These problems with drain side programming dictate high channel
23 currents, care in overerase, potentially high drain voltages,
24 and unfavorable fields (potentially subducting the channel below
25 the surface at the drain side and driving electrons downward
26 away from the floating gate).

27 Consequently, in accordance with the teachings of this
28 invention, programming efficiencies (I_G/I_D) ranging from
29 10^{-5} to 10^{-3} are possible, with I in the range of 1mA during
30 programming, which is two to three orders of magnitude smaller
31 than conventional drain side programming. This offers the
32 potential for very fast system level programming by allowing the
33 programming of 100 times as many memory cells in parallel,

thereby achieving a 100 fold increase in effective program speed compared with prior art drain side programming.

TABLE 1

State Table & Operating Conditions (Fig. 1b)

Node Operation	Poly 3 (Word line)	Poly 2 (Steering Gate)	Drain & (BN Drain)	Source & (BN Source)
R STANDBY	0v	0v	1.0v or 0v	1.0v or 0v
R L READ SELECTED	5v	0v	1.0v or 0v	0v or 1.0v
E A			0v	1.0v
A T READ UNSELECTED	5v	0v	1.0v	1.0v
D E				
D				
E R ERASE UNSELECTED	5v	0v	0v	0v
R E				
A L				
S A				
E T ERASE Option 1	5v	-10to-17v	0v	0v
E or Option 2	12-22v	0v	0v	0v
D				
P R PROGRAM	H1.5v	14-20v	5-7v	0v
R E SELECTED				
O L				
G A				
R T PROGRAM	0v	14-20v	5-7v	0v
A E UNSELECTED	H1.5v	14-20v	5-7v	5-7v
M D	0v	14-20v	0v	0v

A major feature of the cell of this invention is the decoupling of the select function (in this case poly 3 select transistor 110 in Figure 1b) from the steering function (poly 2 control gate 108). During programming, this allows the independent control of cell selection/drain current throttling via poly 3 word line 109 bias (biased at slightly higher than

VT_{p3}) and strong positive voltage coupling onto floating gate 107 (by raising poly 2 control gate 108 to a high voltage, such as about 12 volts). Also, in accordance with the teachings of this invention, the drain voltage can be adjusted independently of steering and select transistor voltage levels, to optimize programming.

During read, the decoupling feature of this invention provides two important advantages, and one exciting side benefit.

1. The ability to set control gate 108 at the optimum voltage level for memory state sensing, i.e. the best balanced reference point for both programmed and erased states. This independence is in contrast to conventional cells wherein the control gate also serves as the select transistor, dictating a voltage level consistent with selection (e.g. Vcc = 5v +_ 10%).

2. Improved margin by virtue of being a fixed, (potentially regulated) reference voltage, eliminating the Vcc variation of +_ 10% inherent to the word line bias levels. (This alone could improve the floating gate memory window by about 0.6v).

3. A side benefit of the ability to independently set the control gate voltage bias discussed above, offers the possibility of a simple way for re-referencing the memory cell for multi-state (i.e. more than conventional 2-state) encoded data. For example if the cell is encoded into three level states, (such as logical 1 = strongly erased/high conducting, logical 2 = partially programmed/ weakly conducting; logical 3 = strongly programmed,) then the control gate voltage can be set at two different levels in a two pass read scheme. For example, in the first pass read the control gate voltage would be set at about 0v to discriminate between the logical 1 state and the logical 2/logical 3 states. In the second pass read the control/gate voltage is set to about 2v, to discriminate between the logical

3 state and the logical 1/logical 2 states. By combining the information of this two pass read (e.g. according to Table 2) the original state of the 3 state cell is recovered. This biasing can be done independently of sense amp reference cell considerations allowing a single sense amp/reference cell circuit to detect the different states via a multi-pass read scheme.

TABLE 2

READ	PASS 1	PASS 2
<u>STATE</u>	<u>[Ref. = 0v]</u>	<u>[Ref. = 2]</u>
1	Hi	Hi
2	Lo	Hi
3	Lo	Lo

The two options for erase operation/bias conditions shown in Table 1 stem from two different sets of considerations. The first option shown brings poly 2 control gate 108 to a large negative voltage, but allows poly 3 word line 109 to remain at a low voltage (e.g. 0v to 5v). This is desirable since the word lines and their decoders are preferably high performance, and repeated many times with a tightly pitched requirement, making high voltage word line requirements more difficult and real estate consuming to implement. Poly 2 control or steering gate 108 on the other hand could be common to a multiplicity of word lines (e.g. a sector consisting of 4 or more word lines), putting less demands on real estate and minimal impact to performance. Possible drawbacks of this approach are the process and device requirements to support both negative as well as positive polarity high voltage circuitry, and reduced steering effectiveness in that the channel cannot assist in steering by virtue of it being held at or near ground (i.e. can't go to large negative potential).

Note that poly 2 is used only as a steering electrode during all three operations. Poly 3, which is the word line connection to the X-decoder, only sees 0V to 5V (other than for erase option 2), and its capacitance can be made relatively small. It is relatively easy to generate +5V and -17V on poly 2 since both writing and erasing are slow operations relative to reading and there is no DC current drain. The -17V does require high voltage PMOS in the erase decode, but the +5V on poly 3 aids in reducing the maximum negative voltage required on poly 2 during erase.

The second option of using high word line voltage bias for erase eliminates both of the above potential drawbacks, but burdens the high performance, tightly pitched word line/driver with high voltage requirement.

Figure 1d is a plan view of one embodiment of an array consisting of a plurality of cells constructed as just described with respect to Figures 1a-1c, and using similar reference numerals. Also shown, are channel stop isolation regions 180.

Figure 1e shows a block diagram of a memory array similar to that shown in the plan view of Figure 1d which is organized by sectors, with appropriate control circuitry. Operation of one embodiment of such a memory array organized by sectors is shown in Figure 1f, where the abbreviations used have the following meanings:

FLT = float

V_{BE} = bit line erase voltage

V_{WE} = word line erase voltage

DI = data in

DIV = data in during verify operation

V_{CEU} = control gate erase voltage - unselected

V_{CE} = control gate erase voltage - selected

S.A. = sense amplifier

V_{CM} = control gate margin voltage (during verify operation)

V_{CP} = control gate program voltage

V_{CR} = control gate read voltage

1 V_{ce} = control gate erase voltage

2

3 As shown in Figures 1e and 1f, in this embodiment sectors are
4 formed by a single column or a group of columns having their
5 control gate connected in common. This allows a high speed shift
6 register to be used in place of a row decoder in order to
7 serially shift in a whole block of column data for the word
8 lines, with the data for each word line being contained in the
9 shift register on completion of its serial loading. The use of
10 such a high speed shift register saves circuit area on an
11 integrated circuit by serving both encoding and latching
12 functions normally performed by a row decoder. Furthermore,
13 speed is improved by including a parallel loaded buffer register
14 which receives data in parallel from the high speed shift
15 register and holds that data during the write operation. While
16 the write operation takes place based upon the data stored in the
17 buffer register, the high speed serial shift register receives
18 the next block of data for subsequent transfer to the buffer
19 register for the next write operation. In one embodiment of this
20 invention, each sector has an associated latch for tagging that
21 sector in preparation for an erase of a plurality of tagged
22 sectors.

23 In one embodiment of this invention, a sector is formed in a
24 group of four cell columns, each column being 1024 bits tall with
25 a common control gate and an associated sector latch. In this
26 embodiment, verification of programming is performed in parallel
27 on all to-be-programmed cells in a single column. Logical 0
28 state cells have word lines at 0 volts while logical 1 state
29 cells have word lines at a positive voltage, such as 5 volts.
30 The control gate and drain voltages are reduced to a verify level
31 to allow for proper margin testing and the bit line current is
32 monitored. If all of the to-be-programmed cells have been
33 properly programmed, the bit line current will be 0 or
34 substantially so. If not, it is known that one or more of the

1 to-be-programmed cells in the column have not been properly
2 programmed, and another write operation is performed on the
3 entire column, thereby assuring that any incompletely ones of the
4 to-be-written cells are again written. An additional verify step
5 is performed to verify that the column has been properly
6 programmed.

7 One embodiment of a process suitable for fabricating the
8 structure having the cross-sectional view of Figure 1b is now
9 described. This embodiment can be implemented in a very small
10 area with no need for an isoplanar oxide when utilizing a virtual
11 ground, allowing an isolation implant to be placed in the
12 remaining field which is not covered by diffusions or
13 polycrystalline silicon and avoids susceptibility to substrate
14 pitting associated with the SAMOS etch in the field isolation
15 region not covered by poly 1. This is achieved, for example,
16 with the following process sequence:

17

18 1. Form BN^+ bit lines in vertical strips. Grow approximately
19 1500Å oxide on top of BN^+ , and approximately 200-300Å gate oxide.

20

21

22 2. As shown in Figs. 7a and 7b, deposit poly 1 to a suitable
23 conductance and etch in horizontal strips perpendicular to the BN^+
24 diffusion. Fill the spaces between adjacent strips of poly 1
25 with deposited oxide, such as CVD followed by an etch back. This
26 approach protects the field isolation regions, and if desired it
27 can be preceded by a boron channel stop implant.

28

29 An alternative for steps 1 and 2 of the above process sequence is
30 forming horizontal strips of isolation oxide first, and then
31 depositing P_1 and etched back in RIE to fill and planarize the
32 horizontal grooves between adjacent strips of isolation oxide.

33

- 1 3. Form thin dielectric 140 such as ONO of approximately 300-400
2 Å. covering poly 1 strips.
3
4 4. Deposit poly 2 and form a suitably thick dielectric overlayer
5 (e.g., approximately 2000-3000 Å of CVD densified oxide). Etch
6 this oxide and underlying poly 2 in long vertical strips parallel
7 to bit line (BN⁺) diffusions.
8
9 5. Form oxide spacers 62 along edges of poly 2 and use edge of
10 these spacers to define the floating gate by etching off exposed
11 poly 1 (i.e. poly 1 not covered by poly 2 or by spacer).
12
13 6. Form tunnel erase oxide in a conventional manner, as
14 described in U.S. patent application serial number 323,779, filed
15 March 15, 1989, over exposed edges of poly 1 as well as gate
16 oxide over the channel of the select transistor (channel 106-1 in
17 Figure 1b).
18
19 7. Deposit poly 3 or polysilicide, and form word lines in
20 horizontal strips.
21

22 Another embodiment for achieving a virtual ground cell
23 without the use of the buried diffusion formed early in the
24 process is now described. In place of the BN⁺ of step 1, after
25 step 6 a photoresist (PR) masked arsenic source/drain implant
26 103a is used, self-aligned to one edge of poly 2 108 after poly
27 1 107 stack formation but leaving an unimplanted region along the
28 other edge to become the poly 3 controlled select transistor
29 channel (see Figure 8). The isolation oxide thickness formed
30 earlier between poly 1 strips is made sufficiently thick to
31 withstand the self-aligned poly 2/1 stack etch without exposing
32 the substrate to pitting, but thin enough such that following
33 this stack etch it is readily removed to expose the substrate to
34 the source drain implant. This offers the benefit of reduced

1 thermal drive of the arsenic junction laterally facilitating
2 scaling. The remainder of the process steps of this embodiment
3 follows the prior embodiment.

4
5 In summary, the novel cell of this invention offers the
6 following benefits.

- 7
- 8 * Very low programming current.
- 9 * Low programming drain voltage requirement/eliminating
10 the need for high voltage.
- 11 * Immunity of Programmability to increased levels of
12 erase.
- 13 * Adjustability of memory state for optimum read of both
14 program and erased states.
- 15 * Improved margin by elimination of sensitivity to $\pm 10\%$
16 Vcc variation on the steering element.
- 17 * Potential for pure low voltage word line/decoder
18 implementation.
- 19 * Facilitates multi-state cell sensing.
- 20 * Reduced susceptibility to source side hot-electron
21 programming induced trapping by establishing a separate
22 threshold control region at the drain.
- 23

24 A second array embodiment is similar to that of Figure 1d but
25 uses the cell embodiment shown in Figure 1b, to form a row
26 oriented sector architecture, is shown in Figure 1g. A sector
27 consists of a group of rows, four in this example, which are
28 erased together. Erase uses option 2 of Table 1, for this row
29 oriented sector architecture, bringing all the poly 3 word lines
30 of a sector to high voltage. The poly 2 steering gate is common
31 to a group of N sectors where N can range from 1 to the full size
32 of the memory array. Similarly the BN+ columns can alternatively
33 continuously span the full length of the array or be broken down
34 into a collection of shorter length, local columns. These

connect to a global (full array length) column through a select transistor driven by an additional level of decoding. The local columns can range from 1 to N sectors. The preferred embodiment is to have local columns span the same number of sectors as the poly 2 steering gate. A preferred number of sectors, N, spanned by local columns and poly 2 steering is around 8. This is because if N is much smaller than 8, the area overhead for local column section devices and poly 2 steering gate routing is high in relation to the area of arrayed cells, while if N is much larger than 8, the benefits of having localized columns and poly 2 steering diminish. These benefits are: (1) reduced bit line capacitance improving read performance; (2) reduced repetitive exposure on unselected sectors to the raised voltage conditions on drains and steering electrodes when programming one sector within the N-sector group, and associated potential disturb phenomena; and (3) increased confinement of array related failures thereby increasing the efficiency of replacing such failures. Read, program and unselected conditions are as described in Table 1, during read or program. The poly 3 word line in the selected row within the selected sector is turned on, 5 volts for read and approximately 1 volt for programming. Concurrently, the drain to source bias conditions are applied to the columns, approximately 5 volts for program and approximately 1.0-1.5 volts for read. In one embodiment, alternate bits in a selected row are programmed simultaneously, thereby permitting all bits in a selected row to be programmed utilizing two programming operations. In a similar manner, in this alternative embodiment, alternate bits in a selected row are read (or verified) simultaneously, thereby permitting all bits in a selected row to be read (or verified) utilizing two read (or verify) operations. After one row in the sector has finished reading or writing, the next row is selected, and so forth to the end of the sector. The resulting row oriented sector architecture and array operation is much more conventional than

the column oriented sector of the first embodiment, and consequently operates in a more traditional manner. Both embodiments share the intrinsic low power capability of this invention, but the row oriented sector embodiment requires, in addition, a full complement of data registers to support massively parallel write and verify features.

Figure 2a shows an alternative array embodiment of this invention which does not utilize buried diffusion regions. Thus, source region 102 and drain region 103 are formed in a conventional manner and not buried by a thick dielectric layer as is the case in the embodiment of Figure 1b. A plurality of memory cells are shown in Figure 2a along a cross section of a typical array structure, with elements of one such cell numbered using reference numerals corresponding to similar structure in Figure 1b. Table 3 depicts an example of the operating conditions appropriate for the embodiment of Figure 2a. This is a more traditional cell approach compared to the buried diffusion cell, with source/drain diffusions formed after all the polycrystalline silicon structures are formed. It requires one drain contact to metal bit line for every 2 cells, making it approximately 30% to 50% larger than the buried diffusion cell with similar layout rules. In all other respects, this alternative embodiment offers the same benefits as listed above for the buried diffusion embodiment of Figure 1b.

Figure 2b is a plan view of one embodiment of an array of memory cells constructed as described above with reference to Figure 2a.

Figure 2c is an equivalent circuit diagram depicting the organization of such a memory array in sectors, with appropriate operating conditions and voltages shown. The preferred embodiment for a sector organized array uses two word lines which straddle a source line as part of a sector, along with their associated poly 2 steering gates and source line. A full sector consists of some multiple of such pairing (e.g. 2 such pairs or

1 4 word lines, each word line containing 128 bytes and overhead
2 cells, and straddling two source lines, constitute one sector).
3 As shown in the embodiment of Figure 2c, the steering lines
4 are connected together within a sector as are the source lines
5 (i.e. a sector which consists of row lines grouped together
6 respectively and driven by common drivers). The embodiment
7 described here confines the write operation to the sector being
8 written to, while the bit line bias conditions (2.5v during read
9 and approximately 5v possible during write) are non-disturbing to
10 the cells because the bias is applied to the select transistor
11 side of the cell and not to the floating gate side. In a two
12 state cell, to write the cell to a logical one, the bit line is
13 held at zero volts, causing the cell to program via source-side
14 injection. Conversely, to inhibit writing, the bit line is held
15 high (typically about 5 volts), thereby cutting off the channel,
16 leaving the cell in the erased state.
17 Sector erase takes place by tagging the selected sector and
18 raising the associated row lines to a sufficiently high voltage
19 to erase the floating gates to their required erased levels.
20 Because of the low programming currents associated with
21 source side injection (approximately 1-5 microamps/cell), massive
22 parallel programming is made practical, e.g. a full row line of
23 approximately 1000 cells is programmed in a single operation with
24 total current less than approximately 1-5mA, thus providing more
25 than 100 times more efficiency than prior art drain side
26 programming arrays.
27

TABLE 3

State Table & Operating Conditions (Fig. 2a)

<u>Node</u> <u>Operation</u>	<u>Poly 3</u> <u>(Word</u> <u>line)</u>	<u>Poly 2</u> <u>(Steering</u> <u>Gate)</u>	<u>Drain</u>	<u>Source</u>
R R STANDBY	0v	0v	Don't care	0v
E E READ SELECTED	5v	0v	2.5v	0v
A L READ UNSELECTED	5v	0v	Don't care	0v
D A				
T				
E				
D				
E R STANDBY	0v	0v	0v	0v
R E				
A L				
S A				
E T ERASE Option 1	12v-22v	0v	0v	0v
E				
D Option 2	5v	-10v to -12v	0v	0v
P R PROGRAM	H1.0v	14-20	0v	5v-8v
R E SELECTED				
O L				
G A				
R T PROGRAM	0v	14-20	0v	5v-8v
A E UNSELECTED	H1.0v	14-20	5v	5v-8v
M D	0v	14-20	5v	5v-8v

Figure 3 is a graph depicting the gate current into poly 1 gate 107 of Fig. 1b (which is not floating in the Figure 3 test device to allow this measurement to be made) as a function of poly 1 gate voltage ($V_{\text{poly 1}}$) while keeping the select transistor 110 V_{p2} at just above its threshold. In this way most of the potential drop in channel 106 of Figure 1 occurs in channel portion 106-1 underneath gate 109 of select transistor 110, and electrons accelerated in this channel are then injected onto floating gate 107. From Fig. 3 it is seen the hot electron programming injection efficiency of this device is phenomenally high.

1 Various embodiments of a process suitable for fabricating a
2 structure in accordance with the embodiment of Figures 1a-1d are
3 now described. Reference can also be made to copending U.S.
4 Application Serial No. 323,779 filed March 15, 1989 (now U.S.
5 Patent 5,070,032), and assigned to SunDisk, the assignee of this
6 invention. Reference may also be made to fabrication process
7 steps described earlier in this application. A starting
8 substrate is used, for example a P type substrate (or a P type
9 well region within an N type substrate). A layer of oxide is
10 formed, followed by a layer of silicon nitride. The layer of
11 silicon nitride is then patterned in order to expose those areas
12 in which N⁺ source and drain regions are to be formed. The N⁺
13 source and drain regions are then formed, for example, by ion
14 implantation of arsenic to a concentration of approximately
15 $1 \times 10^{20} \text{ cm}^{-3}$. The wafer is then oxidized in order to form oxide
16 layers 104 and 105 in order to cause source and drain regions 102
17 and 103 to become "buried". Note that for the embodiment of
18 Figure 2a, this oxidation step is not utilized, as the source and
19 drain regions are not "buried". Rather, the source and drain
20 regions are formed after all polycrystalline silicon layers are
21 formed, in a conventional manner. The remaining portion of the
22 nitride mask is then removed, and the oxide overlying channel
23 regions 106-1 and 106-2 is removed. A new layer of gate oxide
24 overlying channel regions 106-1 and 106-2 is formed, for example
25 to a thickness within the range of 150Å to 300Å and implanted to
26 the desired threshold (e.g. approximately -1v to +1v).
27 Polycrystalline silicon is then formed on the wafer and patterned
28 in order to form floating gate regions 107. If desired, the
29 polycrystalline silicon layer is patterned in horizontal strips
30 (per the orientation of Figure 1a), with its horizontal extent
31 patterned at the same time as the patterning of the second layer
32 of polycrystalline silicon, as will be now described. Following
33 the formation polycrystalline silicon layer 107 at this time, a
34 layer of oxide or oxide/nitride dielectric is formed over the

1 remaining portions of polycrystalline silicon layer 107. A
2 second layer of polycrystalline silicon 108 is then formed and
3 doped to a desired conductivity, for example 30 ohms/square. The
4 second layer of polycrystalline silicon is then patterned into
5 vertical strips (again, per the orientation of Figure 1a). If
6 the horizontal extent of polycrystalline silicon layer 107 was
7 not earlier defined, this pattern step is also used to remove the
8 layer of dielectric between the first and second layers of
9 polycrystalline silicon in those areas where the first layer of
10 polycrystalline silicon is to be patterned simultaneously with
11 the patterning of the second layer of polycrystalline silicon.
12 Following the first layer patterning, an additional layer of
13 dielectric is formed on the wafer to form the gate dielectric
14 above channel region 106-1, and above any other areas in the
15 silicon substrate to which the third layer of polycrystalline
16 silicon is to make a gate. These regions can then be implanted
17 to the desired threshold voltage (e.g. approximately 0.5v to
18 1.5v). The third layer of polycrystalline silicon is for a
19 transistor (ranging from 200Å to 500Å in thickness) then formed
20 and doped to appropriate conductivity, for example 20
21 ohms/square. Polycrystalline silicon layer 109 is then patterned
22 in order to form word line 109.

23 In one embodiment of this invention, polycrystalline silicon
24 layer 107 is patterned to form horizontal stripes and channel
25 stop dopants (e.g. boron) are implanted into the exposed areas
26 therebetween in order to form high threshold channel stop regions
27 between adjacent rows of a memory array. The thickness of the
28 gate dielectric between channel 106-2 and polycrystalline silicon
29 floating gate 107 can range from approximately 150 angstroms or
30 less to approximately 300 angstroms or more, depending on
31 performance tradeoffs. For increased drive for reading, a
32 thinner gate dielectric is desired while for increased coupling
33 between polycrystalline and silicon control gate 108 and floating

gate 107 (helpful during programming) a thicker gate dielectric is desired.

Second Embodiment

Figure 5 is a two-poly embodiment in which programming occurs by taking drain 303 high, for example about 10V while raising control gate 308 just sufficiently so as to turn on select transistor 310. Since this V_{CG} voltage can vary from one device to another it is possible to achieve the optimum injection conditions by keeping V_{CG} at about 3V while raising source (virtual ground) 302 in a sawtooth fashion from about 0 to 3 volts and back to 0 again, with a period on the order approximately 1 microsecond.

This ensures that at some point along the sawtooth the optimum injection conditions are met. Reference can also be made to European Patent Application Serial No. 89312799.3 filed August 12, 1989. To further enhance programming efficiency, in one embodiment a programming efficiency implant 330 (shown in dotted line) is introduced at the source side. To read the device, its source is 0V, drain is approximately 1.0v and V_{CG} approximately 4.5-5v. To erase we employ poly 1-poly 2 tunneling between floating gate 307 in word line 308 at the tunneling zone, consisting of one or more of the floating gate edges, sidewall, corners of the top edge, portions of the top and portions of the bottom, of floating gate 307, associated with a tunnel oxide (400Å-700Å). Erase takes place with V_{CG} approximately 12-22V, $V_D = 0V$, $V_S = 0V$. A capacitive decoupling dielectric (approximately 1500 to 2000Å thick) 340 is formed on top of poly 1 to reduce the capacitance between poly 1 and poly 2.

In one embodiment of this invention, a high electrical field region is created in the channel far away from the reverse field region located in conventional devices near the drain. This is achieved, for example, by utilizing region 330

1 of increased doping concentration at the boundary between
2 channels 306-1 and 306-2 under floating gate 307. In one
3 embodiment, the width of region 330 is on the order of 0.1
4 microns. A larger dimension for region 330 can be
5 counterproductive, reducing the select transistor drive with no
6 gain in efficiency.

7 Figure 4 depicts the electrical field distribution along
8 channels 306-1 and 306-2 in structures with and without P+
9 doped region 330. In a structure without region 330 and
10 improperly biased select transistor the electron injection can
11 take place in the high field region near drain 303. Because of
12 the vertical field reversal region near drain 303, the
13 resultant injection efficiency is reduced. In a structure with
14 region 330 the injection takes place in the high field region
15 located at region 330, far away from the field reversal region.
16 Because of this, increased injection efficiency is achieved.

17 From the processing side there are three problems which
18 must be addressed properly:

- 19
- 20 1. The formation of sufficiently thin/high quality gate
21 dielectric over BN+, which tends to oxidize more quickly than
22 undoped silicon.
 - 23 2. The misalignment between poly 1 and the buried N+ drain
24 diffusion strongly affects the coupling ratios for programming
25 and erase. This can be overcome at the expense of an increase
26 in cell area by not using a virtual ground array, but instead
27 a shared source array.
 - 28 3. This array permits floating gate 307 to completely overlap
29 the buried N+ diffusion in a dedicated source arrangement,
30 eliminating this alignment sensitivity. Unfortunately, this
31 array requires an extra isolation spacing adjacent to the BN+
32 to prevent the poly 1 extension beyond BN+ in the direction
33 away from channel 306-2 to form a transistor in the neighboring
34 cell.

1 To achieve small cell size in the buried diffusion
2 direction a channel stop isolation is used between adjacent
3 cells, plus a self-aligned stacked etch to simultaneously
4 delineate poly 2 and poly 1. This is difficult to do without
5 pitting the substrate as well as the exposed BN+ when etching
6 the exposed poly 1 between adjacent cells. This is especially
7 difficult to avoid when etching the decoupling oxide (1500 -
8 2000Å thick on top of poly 1 in order to expose poly 1, since
9 the substrate unprotected by poly 1 also becomes exposed, so
10 that when poly 1 is etched, the substrate in those regions
11 becomes pitted.

12 This will therefore require formation of a thick dielectric
13 region as part of the field isolation process protecting the
14 substrate in the space between the poly 2 word lines. This can
15 be accomplished by using a process as described in U.S. Patent
16 Application Serial No. 323,779, filed March 15, 1989, and
17 assigned to SunDisk, the assignee of this application. This is
18 actually forming trench isolation, but with BN+ abutting this
19 trench, we may experience severe junction leakage as well as
20 loss of a portion of the BN+ conductor. This cell of this
21 second embodiment is attractive because it is double poly, low
22 programming current, very fast programming, programming away
23 from drain junction, small and scalable cell. Cell size is
24 quite attractive as indicated below for three representative
25 geometries:

26 1.0m geometries: cell = $4.0 \times 2.0 = 8.0\text{m}^2$
27 0.8m geometries: cell = $3.2 \times 1.6 = 5.2\text{m}^2$
28 0.6m geometries: cell = $2.3 \times 1.2 = 2.8\text{m}^2$

29 30 Third Embodiment

31 Figure 6 is a cross-sectional view of alternative
32 embodiment of a two poly cell, using source side injection for
33 programming, aided by strong coupling to buried N+ drain 403,
34 which acts also as a second control gate. Erase is by Fowler-

1 Nordheim tunneling to channel 406 through a small thinned oxide
2 region, formed for example to a thickness of about 100\AA , by
3 utilizing a thin polyspacer. These process steps would be as
4 follows: Once the drain oxide is formed (i.e. the oxide above
5 drain 403), a first layer of poly, (approximately 2000\AA to
6 4000\AA thick) is deposited and a thin nitride dielectric is
7 deposited on top. These layers are then etched using a poly 1
8 mask to delineate the lateral extent (as shown in Figure 6) of
9 the poly 1. A second layer of nitride is then deposited and
10 anisotropically etched back to underlying oxide, leaving the
11 initial nitride layer on top of poly 1 plus nitride spacers
12 along the poly 1 sidewalls. This protects the poly 1 sidewall
13 from subsequent oxidation, allowing electrical contact to be
14 made as later described. The exposed oxide layer over the
15 channel portion of the substrate is then stripped and regrown
16 to the 100\AA thickness required for tunneling, while a
17 photoresist masked pattern protects oxide over the exposed, BN^+
18 side of the poly 1 from being stripped. The nitride layers
19 surrounding poly 1 prevent oxide from forming on that poly.
20 The thin nitride is then etched off using a highly selective
21 etch which does not attack or degrade the 100\AA tunnel oxide
22 (e.g. hot phosphoric or plasma etch). This is followed by a
23 second poly deposition which electrically contacts the first
24 poly on its top surface and its sidewalls. This structure is
25 then etched using an anisotropic poly-silicon etch, with etch
26 being terminated with the re-exposure of the oxide layers over
27 substrate beneath the second deposited poly layer. This
28 completes the formation of the poly 1 floating gate stripe
29 shown in Figure 6. The remaining process is similar to that of
30 the second embodiment.

31 In this embodiment, programming is from hot channel
32 electrons injected from grounded source diffusion 402 with
33 drain 403 held at about +8v and fixed control gate of around
34 1.5v. Alternatively, programming is performed by hot channel

1 electrons from source diffusion 402 utilizing a sawtooth
2 control gate voltage ranging from 0 volts to a peak voltage
3 approximately 3 volts, as described previously for the second
4 embodiment. Read is achieved with $V_{DS} = 1.5V$, $V_s = 0$, $V_C =$
5 $+5V$. Erase is achieved with $V_{CG} = -22V$, $V_s = V_d = 0V$. In this
6 embodiment, the poly 2 word line 408 will carry the +5 volts
7 during read and the -22 volts during erase, thereby requiring
8 an X-decoder capable of serving this purpose. Coupling
9 considerations require that $C_{P2P1} > C_{P1D}$, which is unfavorable
10 for programming. Therefore the cell must be optimized for
11 balancing erase against programming by adjusting oxide
12 thicknesses and floating gate threshold to the optimum point.
13 There is less of a problem with pitting the field regions
14 between cells in the poly 1 direction (because poly 1--poly 2
15 oxide or ONO is thin). This may obviate the need for the
16 additional thick oxide field region described for the second
17 embodiment. However, there is the additional process
18 complexity of forming the thin oxide region and extra space
19 needed to place this thin oxide region sufficiently far from
20 the source diffusion.

21

22 Alternative Operating Methods

23 A number of alternative methods are possible to program the
24 source side injection cells described in the previous
25 embodiments. Strong capacitive coupling (for example, using
26 thin ONO) is required in the second and third embodiments
27 between poly 2 and drain, and between poly 2 and poly 1,
28 respectively, for programming. During operation, one embodiment
29 applies V_d at 5 to 7v, $V_s = 0$, the control gate voltage V_{CG} is
30 raised to just turn on the control gate channel, and V_{p2} is on
31 the order of about 12 volts or more. Alternatively, the source
32 body effect is used to advantage. In this alternative
33 embodiment, rather than bringing control gate to a specified
34 value to just turn on the channel, the control gate is brought

to a value greater than the voltage required to just turn on the channel (e.g., approximately one volt above) and a pull-down circuit is used (e.g., a high impedance resistor or a current sink) for providing approximately 1 μ A current flow via source debiasing. Alternatively, the control gate voltage V_{CG} can be operated in a sawtooth fashion from between 0 volts to about +3 volts, as mentioned previously with respect to European patent application serial number 89312799.3.

Multi-bit Cells

In an alternative embodiment of this invention, such as is shown in Figures 9a and 9b, a novel structure is taught including a multi-bit split gate cell, using source side injection programming and using poly-to-poly tunneling for erase. The following describes, in more detail, the operation of one embodiment of such a structure of this invention.

Basic read operation for such a cell consists of applying appropriate control gate bias (e.g. 8v - see TABLE 4) to the unread portion (henceforth for convenience to be termed the transfer portion), while applying the required read control gate bias to the portion being sensed (in multi-state this would be a bias level appropriate to the state being sensed for). In one embodiment, the select gate bias is held at approximately 1.5 volts to keep total cell current limited (e.g. to about 1 microamp), independent of the floating gate conduction level. Alternatively, the select gate bias is maintained at any desired level, e.g. about 5 volts, depending on the current sensing requirements. Similarly, to program a bypass applied on the transfer portion (about 12v) and a writing potential on the control gate portion (again in multi-state this would be a bias level appropriate to the state being written), with the select gate bias throttled for source side emission (about 1.5v), and the drain bit line (the bit line adjacent the to-be-programmed floating gate) raised to about 5v

for programming, with the source bit line (adjacent to transfer portion) grounded.

TABLE 4
OPERATING MODES/CONDITIONS

CONDITION	BL1	CGL2	SG1	CGR2	
BL2					
R STANDBY	X	X	0v	X	X
E READ UNSELECTED	FLOAT	X	1.5v	X	FLOAT
A READ FGL12	0v	READ VREF	1.5v	8v	1.5v
D READ FGR12	1.5v	8v	1.5v	READ VREF	0v
ERASE	0v	0v	VE	0v	0v
P					
R STANDBY	X	X	0v	X	X
O PROG UNSELECTED	FLOAT	X	1.5v	X	FLOAT
G PROG FGL12	5v	PROG VREF	1.5v	12v	0v
R PROG FGR12	0v	12v	1.5v	PROG VREF	5v
A					
M					

NOTES: X - DON'T CARE; VE - OPTIMUM ERASE VOLTAGE ($\sim < 20v$)

Following are some key advantages of the multi-bit cell of this embodiment of this invention:

- (1) Approaches $(2 \cdot \lambda)^2$ cell size
- (2) Highly self-aligned
- (3) High efficiency source side programming, resulting in lower power and lower voltage requirements, allowing greater parallelism during write
- (4) Attractive for scalability
- (5) Totally immune to overerase

This cell can achieve $(2 \cdot \lambda)^2$ cell size, where λ is the minimum lithographic feature, because (1) each of its

1 lateral component parts, in both its word line and bit line
2 directions, can be formed using this minimum lambda feature,
3 and (2) the various critical components are self-aligned to one
4 another, obviating the need to increase cell size to
5 accommodate lithographic overlay registration requirements.
6 For example, viewing along the row or word line direction, the
7 floating gate poly2/1 self-aligned stacks and their underlying
8 channels can be formed using the minimum feature lithographic
9 width (lambda), while the transfer channels and bit line
10 diffusions can be simultaneously delineated using the minimum
11 lithographic space between features (also lambda), giving it a
12 $(2*\lambda)$ minimum pitch capability along this direction.
13 Similarly, looking along the poly2 steering gate in the bit
14 line direction, the channel regions underlying poly1 floating
15 gate and poly3 word line can be formed using the minimum
16 lithographic feature (lambda), while the isolation region
17 between word line channels can be formed by the minimum
18 lithographic space (also lambda), again achieving the minimum
19 pitch of $(2*\lambda)$. In this way, the cell achieves the
20 $(2*\lambda)^2$ minimum layout area. It is in fact a self-aligned
21 cross-point cell, the poly2/1 stack and corresponding channel
22 being fully self aligned to the transfer channel and bit line
23 diffusions, and in the orthogonal direction the isolation being
24 self-aligned to the channel areas. When combining this with
25 the low voltage requirement made possible by the source-side
26 injection programming mechanism, this makes it an ideal element
27 for still further scaling (i.e. smaller lambda). Finally, its
28 immunity to overerase comes from the following two factors:
29 (1) the presence of the series transistor channel select
30 region, which fully cuts off cell conduction when deselected,
31 independent of degree of erasure, and (2) the source-side
32 injection mechanism itself, which is enhanced with strong
33 overerase, in contrast to the more conventional drain-side

1 programming, which becomes retarded by strong levels of
2 erasure.

3 In one embodiment, rather than the use of 100Å tunneling
4 oxide for the erase operation as in the prior art Ma approach,
5 a thick oxide, geometrically enhanced, poly-to-poly tunneling
6 approach is used, as shown for example in Figures 9a and 9b,
7 where the poly3 word line serves the dual function of cell
8 selection and erase anode (one of the architecture/operational
9 approaches taught in the above-mentioned SunDisk Patent No.
10 5,313,421). Figures 10a and 10b shows the equivalent circuit
11 of this cell/array and TABLE 4 summarizes its operation.

12 The advantages of this embodiment include:

13 * Erase unit to follow row line(s), resulting in row oriented
14 sectoring;

15 * Avoids need to use negative voltages, erase being
16 accomplished by holding all electrodes at ground, except for
17 the selected sector(s) poly3 word lines, which are raised to
18 erase potential (about or less than 20v);

19 * High reliability inherent to thick oxide tunneling
20 implementation; and

21 * Improved scalability inherent to the use of the thick
22 interpoly oxide (and consequent reduced parasitic capacitance,
23 both because of the greater thickness and because of the small
24 sidewall vicinity limited tunneling area), combined with the
25 high degree of vertical integration (vertically stacked poly3
26 word line serving the dual role of select gate and erase
27 electrode).

28 Such a cell approach offers the potential for a physically
29 minimal ($4 \times \lambda^2$), highly self aligned, crosspoint cell,
30 which is both very reliable (use of thick oxides and no high
31 voltage junction requirements within memory array), and readily
32 scalable (via the source side injection element and its reduced
33 voltage and more relaxed process control requirements, combined
34 with the inherent scalability of the vertically integrated,

thick oxide interpoly erase element). From a physical point of view therefore, a Gigabit (or greater) density level embodiment based on a 0.25μ technology, has a per bit area of approximately $0.25\mu^2$.

Despite the series nature of the dual gate cell, a four level multi-state (two logical bits per floating gate, or four logical bits per dual gate cell) can be implemented. The key requirement is that the most heavily programmed state plus bias level of the transfer floating gate's control gate be optimally selected to expose the full multi-state conduction range of the memory floating portion, without introducing read disturb. Based on the above example, a four-level multi-state implementation would give a per bit area approaching $0.1\mu^2$ (approximately $0.125\mu^2$).

In summary, the above described dual-gate cell based on the thick oxide, row oriented erase approach offers a novel, non-obvious implementation, one that offers significant improvements over the prior art in scalability, reliability and performance.

Alternative Embodiment Utilizing Negative Steering Cell Operation

The control gate (or steering) bias voltage level or range of levels for reading constitute a powerful parameter in setting the memory window voltage position and corresponding ranges for the steering element during programming operations and the poly3 control/erase element during erase. By allowing this level or range of levels to go below 0v, this allows shifting up of the floating gate voltage memory window (due to its associated charge) by a proportional amount, governed by the steering gate coupling ratio. The net result is the maximum steering gate voltage level, for both sensing and programming, is reduced by that negatively shifted amount.

1 Similarly, with the steering gate taken below 0v during erase,
2 the maximum erase voltage is also lowered, the amount of which
3 is proportional to the steering gate coupling ratio.

4 An important parameter in determining steering voltage
5 magnitudes is the steering gate coupling ratio, RCG (or $R21$) =
6 $C21/CTOT$, where $C21$ is the capacitance between the poly1
7 floating gate and the poly2 steering gate, and $CTOT$ is the
8 total floating gate capacitance. For example, if the net
9 requirement for read plus programming is to capacitively shift
10 the floating gate potential by 10v, then given an RCG of 50%,
11 the steering voltage swing must be scaled up by $1/RCG$, giving
12 a 20v swing. If, on the other hand, RCG is increased to 66.7%,
13 the steering voltage swing drops to 15v, a savings of 5v.
14 Using this 66.7% value, if the read steering bias voltage level
15 (or range) is lowered by 7.5v, the poly3 erase voltage is
16 lowered by $RCG \times 7.5v$, a savings of 5v over the non-lowered bias
17 situation.

18 In order to implement negative steering into an N channel
19 based, grounded substrate memory array, one embodiment utilizes
20 P channel circuitry, capable of going negative of ground, to
21 generate and distribute this bias. In order to support the
22 full steering voltage dynamic range, the N well for such
23 P channel circuitry is biased to the maximum required positive
24 voltage, and the P channel circuitry can thus feed any
25 potential from that value on down to the most negative required
26 (independent of memory array ground). The positive and
27 negative voltage limits are provided from either external
28 supplies or readily generated on chip (for example by N channel
29 based charge pumps for positive bias and P channel for negative
30 bias), since no DC current is required for steering (only
31 capacitive load charging).

32 In one embodiment, a full column oriented array
33 segmentation is implemented to form one sector or a group of
34 row oriented sectors, wherein one sector is read or programmed

1 at any given time. All cells in one sector are erased
2 simultaneously, and one or more sectors can be selected for
3 simultaneous erasure. Column based segmentation breaks a full
4 array into a multiplicity of segmented sub-arrays, thereby
5 eliminating large and/or cumulative parasitics such as
6 capacitance and leakage. Each sub-array has its own set of
7 local bit line diffusions and poly2 steering lines, which are
8 selectively connected by segment select transistor matrixes to
9 corresponding global bit lines and steering lines.

10 Figure 10c exemplifies such a segmentation embodiment,
11 depicting one segment, denoted as SEGMENT I, consisting of N
12 rows of cells (e.g. N equalling 32). For example, each row
13 forms one sector consisting of 2048 dual gate cells or
14 equivalently 4096 floating gate storage elements.
15 Alternatively, a sector can be formed by a group of two or more
16 rows. The long, continuous, global bit lines (typically run in
17 metal) BLk are selectively connected to the local segment
18 subcolumns through the Segment Bit Line Transfer Select
19 transistors 1001, 1002, driven by the SEGi lines. Similarly,
20 the long, continuous global steering lines (typically run in
21 metal) Sk are selectively connected to the local segment
22 steering gates through the Steering Drive Transfer Select
23 transistors 2001, 2002, driven by the STD_ODDi and STD_EVENi
24 lines. In this way array segments are isolated from one
25 another, eliminating the large cumulative parasitics of leakage
26 and capacitance, and providing column associated defect and
27 repetitive disturb confinement.

28 Performance can be increased by simultaneously operating on
29 as many cells in one row as possible (where a row may have
30 anywhere from 1K to 4K floating gate memory transistors),
31 thereby maximizing parallelism. Peak power is not a limitation
32 in such implementation, because of the low cell operating
33 currents inherent to this cell approach both during read and
34 programming operations. Consequently, the number of floating

1 gate transistors per row which can be simultaneously operated
2 on is limited only by addressing constraints and segment decode
3 restrictions. For the embodiment shown in Figure 10c, this
4 allows every 4th floating gate to be addressed and operated on,
5 simultaneously, as outlined in TABLE 5, allowing the full row
6 to be addressed and operated on in four passes as follows.

7 During each pass, two adjacent diffusions are driven to
8 drain potential followed by two adjacent diffusions driven to
9 ground, with that bias pattern repeated across the entire row
10 of cells. In this way global drain/source bias is applied in
11 mirrored fashion to every other of the selected cells,
12 resulting in floating gate bias conditions of odd selected
13 cells being reversely applied to those of the even selected
14 cells. Appropriate biases are placed on the global steering
15 lines, as exemplified in TABLE 5, to satisfy the operation of
16 the targeted floating gates as given in TABLE 4, while the
17 local steering lines of the unselected cells are discharged and
18 left isolated from the global steering lines. Once done, the
19 bias conditions for both global bit/ground lines and
20 targeted/untargeted floating gate steering lines are
21 correspondingly interchanged to operate on the other of the
22 floating gate pair within the selected cells. Once this is
23 completed, similar operation is repeated to the alternate set
24 (i.e. previously unselected set) of cells, thereby completing
25 full row programming in four passes.
26

TABLE 5

		GLOBAL BIT LINES							
		BLK-3	BLK-2	BLK-1	BLK	BLK+1	BLK+2	BLK+3	BLK+4
CELLS									
	K-3L K-3R K-2L K-2R								
	K-1R K-1L KR KL								
	K+1L K+1R K+2L K+2R								
	K+3R K+3L K+4R K+4L								
READ									
PASS 1	SEL UNSEL UNSEL UNSEL	0	1.5	1.5	0	0	1.5	1.5	0
PASS 2	UNSEL UNSEL UNSEL UNSEL	1.5	0	0	1.5	1.5	0	0	1.5
PASS 3	UNSEL UNSEL SEL UNSEL	0	0	1.5	1.5	0	0	1.5	1.5
PASS 4	UNSEL UNSEL UNSEL SEL	1.5	1.5	0	0	1.5	1.5	0	0
PROGRAM									
PASS 1	SEL UNSEL UNSEL UNSEL	5	0	0	5	5	0	0	5
PASS 2	UNSEL UNSEL UNSEL UNSEL	0	5	5	0	0	5	5	0
PASS 3	UNSEL UNSEL SEL UNSEL	5	5	0	0	5	5	0	0
PASS 4	UNSEL UNSEL UNSEL SEL	0	0	5	5	0	0	5	5
ERASE	SEL SEL SEL SEL	0	0	0	0	0	0	0	0

GLOBAL STEERING LINES									
		SK-3	SK-2	SK-1	SK	SK+1	SK+2	SK+3	SK+4
CELLS									
	K-3L K-3R K-2L K-2R								
	K-1R K-1L KR KL								
	K+1L K+1R K+2L K+2R								
	K+3R K+3L K+4R K+4L								
READ									
PASS 1	SEL UNSEL UNSEL UNSEL	VREFR 8	8 VREFR	8 VREFR	VREFR 8	VREFR 8	8 VREFR	8 VREFR	VREFR 8
PASS 2	UNSEL SEL UNSEL UNSEL	VREFR 8	VREFR 8	VREFR 8	8 VREFR	8 VREFR	8 VREFR	8 VREFR	8 VREFR
PASS 3	UNSEL UNSEL SEL UNSEL	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8
PASS 4	UNSEL UNSEL UNSEL SEL	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8	VREFR 8
PROGRAM									
PASS 1	SEL UNSEL UNSEL UNSEL	VREFP 12	12 VREFP	12 VREFP	VREFP 12	VREFP 12	12 VREFP	12 VREFP	VREFP 12
PASS 2	UNSEL SEL UNSEL UNSEL	VREFP 12	VREFP 12	VREFP 12	12 VREFP	12 VREFP	12 VREFP	12 VREFP	12 VREFP
PASS 3	UNSEL UNSEL SEL UNSEL	VREFP 12	VREFP 12	VREFP 12	12 VREFP	12 VREFP	12 VREFP	12 VREFP	12 VREFP
PASS 4	UNSEL UNSEL UNSEL SEL	VREFP 12	VREFP 12	VREFP 12	VREFP 12	VREFP 12	VREFP 12	VREFP 12	VREFP 12
ERASE	SEL SEL SEL SEL	0	0	0	0	0	0	0	0

		SEGMENT _I LINES				ROW LINES	
		STD EVEN_I	STD ODD_I	SEG_I	SELECTED ROW _J LINE	UNSELECTED ROWS	
CELLS							
	K-3L K-3R K-2L K-2R						
	K-1R K-1L KR KL						
	K+1L K+1R K+2L K+2R						
	K+3R K+3L K+4R K+4L						
READ							
PASS 1	SEL UNSEL UNSEL UNSEL UNSEL	0	10	5	1.5	0	
PASS 2	UNSEL UNSEL UNSEL UNSEL UNSEL	0	10	5	1.5	0	
PASS 3	UNSEL UNSEL SEL UNSEL UNSEL	10	0	5	1.5	0	
PASS 4	UNSEL UNSEL UNSEL UNSEL UNSEL	10	0	5	1.5	0	
PROGRAM							
PASS 1	SEL UNSEL UNSEL UNSEL UNSEL	0	14	8	1.5	0	
PASS 2	UNSEL UNSEL UNSEL UNSEL UNSEL	0	14	8	1.5	0	
PASS 3	UNSEL UNSEL UNSEL UNSEL UNSEL	14	0	8	1.5	0	
PASS 4	UNSEL UNSEL UNSEL UNSEL UNSEL	14	0	8	1.5	0	
ERASE	SEL SEL SEL SEL SEL	5	5	5	<20	1	0

1

2 To give an idea of the high speed of this approach with
3 respect to programming, assuming a physical row of 4096 floating
4 gate elements, and $10\mu\text{sec}$ per pass for cell programming, this
5 gives an effective programming time of $\sim 10\text{nsec/bit}$ or a raw
6 programming rate of 4096bits per $40\mu\text{sec}$ (i.e. per $4*10\mu\text{sec}$) or
7 $\sim 12.5\text{MBytes/sec}$.

8 In order to accommodate the negatively shifted steering in
9 this embodiment, the steering segmentation transistor matrix is
10 implemented in positively biased N well, P channel based
11 circuitry.

12 As indicated above, in order to reduce maximum voltage levels
13 required, it is desirable to keep the steering gate coupling
14 ratio relatively high, for example, greater than approximately
15 60%, (see Figure 10a for one embodiment of a cell equivalent
16 circuit). In one embodiment, ONO interpoly2/1 dielectric (with,
17 for example, an effective tox of 200\AA) is used, combined with a
18 cell structure and process approach (described below), which
19 reduces the parasitic substrate and interpoly3/1 capacitances.

20 Parasitic capacitances to substrate and drain are, in one
21 embodiment, kept small by using a narrow channel structure,
22 bounded by much thicker field oxide regions (such isolation
23 structure is described in U.S. patent 5,343,063). By way of
24 example, a cell with a narrow (for example, about 0.1μ wide),
25 approximately 300\AA thick gate oxide channel region bounded by
26 about 1500\AA thick field regions, whose floating gates are laid
27 out so as to substantially overlap those thick field regions (for
28 example with a total overlap of about 0.3μ), would, in
29 combination with the scaled ONO interpoly2/1 , provide steering
30 capacitance magnitudes of around five times larger than those of
31 the floating gate to substrate/drain.

32 In order to reduce the interpoly3/1 capacitance, it must
33 first be noted that in this dual floating gate Flash cell, poly3
34 crosses two edges of the poly1 floating gate, resulting in

1 approximately double the interpoly3/1 capacitance of cells in
2 which poly3 crosses only a single poly1 edge (for which parasitic
3 coupling ratios are typically around 15%). Although the double
4 edge structure may offer benefits to the erase tunneling element
5 (e.g. voltage levels and distributions), its benefit is
6 outweighed by the higher erasing and programming gate voltages
7 needed to offset the associated poorer coupling efficiencies.
8 Therefore, it is desirable to eliminate the capacitive impact of
9 one of these two edges, even if in doing so its erase tunneling
10 contribution is also eliminated. The following discussion
11 describes one embodiment of a process to accomplish this,
12 integrated into the self-aligned diffusion (BN+) formation
13 process.

14 To realize a self-aligned BN+ cell, the BN+ sources/drains
15 must be formed after the poly2/1 stack etch (i.e. self-aligned to
16 poly2) thereby realizing the physically smallest cell. The
17 challenge here is to remove the field oxide locally over the S/D
18 region to allow BN+ As implant, while at the same time preserving
19 sufficiently thick dielectrics surrounding the poly2 steering
20 line for poly3 to poly2 high voltage isolation. The following
21 section details the above mentioned exemplary process.

22 In looking at the twin cell in cross-section (see Figures
23 11a-11e for top view and various cross-sections, and in
24 particular Figure 11e), the process strategy to achieve both
25 self-aligned BN+ formation and poly3/1 coupling reduction lies in
26 the ability to separately process the two distinct regions,
27 namely (1) the vertical strip regions associated with the BN+ and
28 (2) the vertical strip containing the select channel portions.
29 In so doing, the poly3/1 tunneling edge can be restricted to only
30 form adjacent to the select strip, while completely eliminating
31 its formation along the poly1 edge bordering the BN+ strip.

32 This is accomplished in the following manner (refer to Figure
33 12a for cross-sections in row line direction, following some of
34 the key process steps. NOTE: the poly3 row lines are defined

1 here to run horizontally, and the BN+ columns to run vertically).
2 By way of example, the following discussion includes
3 representative numbers for dimensions and thicknesses, assuming
4 a 0.25μ technology (printing minimum lithographic feature size,
5 both width and space, to achieve minimum pitch).

6 Form field oxide 1100 to a thickness of about 1500\AA , and etch
7 it into horizontal strips, adding appropriate channel/field
8 implants prior to or at this step. Use an oxide spacer approach
9 to reduce channel width (for example, reduce from about 0.25μ as
10 etched to about 0.1μ post spacer formation, thereby improving
11 control gate coupling). Grow floating gate oxide 1101,
12 (approximately 300\AA gate oxide). As shown in Sundisk U.S. patent
13 5,343,063, the fabrication steps up through the forming of poly1
14 1102 to a thickness of about 1500\AA are performed. Poly1 is then
15 etched into horizontal strips overlying the channel regions plus
16 generous overlap on the field region to either side of the
17 channel. As with channel width, a spacer approach can be used to
18 decrease the etched poly1 spacing, thereby increasing net poly1
19 overlap of field oxide, or "wings". For example, after the
20 spacer step, poly1 spacing is reduced to about 0.1μ , giving poly1
21 wings of about 0.15μ per side - refer to Figure 11b showing a
22 cross-section through the channel along the column direction for
23 an example of poly1 wings over field oxide. Note that because of
24 the narrow channel widths vis a vis the poly1 thickness, poly1
25 1102 will completely fill the trench, resulting in a
26 substantially planar surface. Next form thin ONO 1103 (for
27 example, having about 200\AA tox effective) on top of and along
28 edges of the poly1 strips. In an alternative embodiment, a
29 portion of the top film is formed as part of an initially
30 deposited poly1 layer stack.

31 Referring to Figures 12a(i) and 12a(ii), deposit a sandwich
32 layer of poly2 1104 (about 1500\AA), thick poly3/2 isolation oxide
33 1105 (approximately 2000\AA), plus a sufficiently thick etch
34 stopping layer 1106 (to block underlying oxide removal when

1 exposed to oxide type etch), and top oxide layer 1107. Using a
2 patterned photoresist masking layer 1108, these are then etched
3 in strips along the column direction, down to the poly1 layer, to
4 form poly2 the steering gate lines. These exposed poly1 regions
5 are overlying the areas to become select channel and BN+.

6 Referring to Figures 12b(i) and 12b(ii), strip previous
7 photoresist and pattern new photoresist layer 1109 to cover and
8 protect exposed poly1 over the select channel regions. Etch
9 exposed poly1 1102 and all its underlying oxide 1101 which cover
10 the to-be-formed BN+ regions. Oxide layer 1107 over etch
11 stopping layer 1106 is used to protect etch stopping layer 1106
12 from being etched by the poly etch as poly1 1102 is being
13 removed. Etch stopping layer 1106 (e.g. thin undoped polysilicon
14 or possibly nitride - must have low etch rate compared to oxide
15 etch rates) is used to prevent that portion of thick poly3/2
16 isolation oxide 1105 not covered by photoresist 1109 from being
17 etched down as oxide 1101 beneath poly1 1102 is etched away. The
18 oxide etch system used is both highly anisotropic (e.g. RIE) and
19 selective vis a vis the underlying silicon substrate, resulting
20 in negligible etching of that substrate, accommodating the large
21 differences in oxide thicknesses being removed between field
22 oxide (approximately 1500Å) and gate oxide regions (approximately
23 300Å). Following completion of all etching, photoresist 1109 is
24 removed.

25 Referring to Figures 12c(i) and 12c(ii), at this point, an
26 option is to implant and drive a sufficient Boron dose to form a
27 p+ DMOS type doping profile adjacent to the BN+ junction
28 (alternatively, this is the point at which the arsenic BN+ is
29 implanted, but the resulting lateral diffusion makes the floating
30 gate channel unnecessarily short). As shown in Figs. 12c(i) and
31 12c(ii), oxide is formed and reactive ion etched back down to
32 silicon to form sidewall spacers 1110 (about 750Å thick, with the
33 thickness here being determined by interpoly3/2 erase high

1 voltage isolation requirements, for example, about 25v). The
2 arsenic BN+ strips are then implanted.

3 Referring to Figures 12d(i) and 12d(ii), a new patterned
4 photoresist layer 1111 is added to cover and protect BN+ strips.
5 The exposed poly1 over channel strips is etched, to expose the
6 selected channel regions. (Since some of the poly1 overlies
7 channel regions and is therefore thicker, while other portions
8 overlie field oxide and is thinner, the same considerations for
9 oxide etch selectivity apply as above, in the oxide over BN+
10 etching case.)

11 Referring to Figures 12e(i) and 12e(ii), once photoresist
12 1111 is stripped, oxide is formed (e.g. via thermal oxidation or
13 some composite oxide) to simultaneously form the poly1 sidewall
14 1112 and corner interpoly3/1 tunneling oxides 1113 (for example,
15 about 350Å), the poly3 gate oxide 1114 over the select channel
16 and oxide 1115 over BN+ (e.g. less than about 300Å - the
17 requirement for both of these oxides being they must be
18 sufficiently thick to reliably hold up to the erase voltage to
19 substrate differential). A select transistor threshold adjust
20 implant can be optionally introduced at this time (e.g.
21 increasing channel dopant concentration to raise select V_t , or
22 introducing compensation implant to reduce select V_t).

23 Referring to Figures 12f(i) and 12f(ii), after deposition and
24 patterning of poly3 (which in one embodiment is polysilicide in
25 order to reduce word line delay) the basic dual gate cell
26 structure is complete. In one embodiment of this invention, a
27 high electrical field region is enhanced in the channel far away
28 from the reverse field region located in conventional devices
29 near the drain and source regions. This is achieved, for
30 example, by utilizing regions 1200 of increased doping
31 concentration at the boundary between the channels 1201 and 1202
32 and transfer channel region 1203. In one embodiment, the width
33 of region 1200 is on the order of 0.1 microns.

1 Using the above dimension and film thickness example values,
2 the total floating gate capacitance becomes about 0.4
3 femtoFarads, and coupling ratios are approximately: Steering
4 Gate (R21) 70%; Erase Gate 20%; Floating gate to Substrate &
5 Drain 10%. Although this R21 value may vary from this figure
6 somewhat in that fringing fields from the other terminals are not
7 accounted for, this approximation indicates adequate coupling
8 ratios are achieved in the dual gate cell, even under aggressive
9 cell scaling.

10 A process variant of the above approach, which can reduce
11 further still the erase coupling, is to completely fill the
12 region over BN+ with an oxide, after BN+ formation. This is
13 done, for example, by depositing a sufficiently thick,
14 undensified (and hence easily etched away compared to underlying
15 densified oxide films) oxide layer, patterning photoresist strips
16 over the BN+ to protect it from etching, and etching away the
17 exposed, undensified film over the select channel strips.
18 Following this step and resist removal, the poly3/1 tunnel oxide
19 process proceeds as outlined above, during which time the oxide
20 filler over BN+ is densified.

21 The above approach and its variant outlines one of a number
22 of possible ways to implement the above described dual floating
23 gate cell into the desired array.

24 In summary, several concepts have been introduced to
25 implementing the TWIN FG cell.

26 Fundamental to the cell is its low power source side
27 programming mechanism, and low power row oriented poly-to-poly
28 erase element. Additionally, its independent steering and
29 selection functions, facilitates low power, multi-state read and
30 programming operations.

31 ONO interpoly2/1 is readily integrated to provide a high
32 capacitive coupling, ultra-low leakage steering element. One
33 embodiment uses a full column segment confinement architecture to
34 substantially reduce parasitic bit line capacitance and leakage.

1 A negatively shifted voltage steering implementation allows
2 reduction of maximum voltage ceilings for both the poly2 steering
3 lines during programming and the poly3 word/erase lines during
4 erase. Under such implementation, one preferred embodiment for
5 the column segmented array architecture is via an N-well isolated
6 P channel steering selection matrix.

7 High steering ratio is achieved by the narrow channel plus
8 field oxide approach to allow formation of wings. A preferred
9 embodiment is described which reduces the interpoly3/1 parasitic
10 as part of a self-aligned BN+ formation process. This replaces
11 the thinner tunneling oxide adjacent one of the two potential
12 tunneling edges with a much thicker isolation oxide. Based on
13 the example used, this approach can give a cell with steering
14 coupling ratio approaching 70%, and parasitic erase coupling down
15 to 20%. Furthermore, based on that example, which uses a 0.25μ
16 technology for the $4 \times \lambda^2$ dual floating gate, poly3 word/erase
17 line cell (where λ is the minimum technology feature size),
18 a physical cell area of $0.25\mu^2$ is realizable, which for 8 (16)
19 level of multi-state translates to an effective cell size
20 approaching $\sim 0.08\mu^2$ ($\sim 0.06\mu^2$) per logical bit. These small sizes,
21 around 100 times smaller than physical sizes of cells used in the
22 4MEG and 8MEG generation of Flash memories, are suitable for
23 building Gigabit density level Flash memories with comparable die
24 sizes and at comparable cost per die.

25 All publications and patent applications mentioned in this
26 specification are herein incorporated by reference to the same
27 extent as if each individual publication or patent application
28 was specifically and individually indicated to be incorporated by
29 reference.

30 The invention now being fully described, it will be apparent
31 to one of ordinary skill in the art that many changes and
32 modifications can be made thereto without departing from the
33 spirit or scope of the appended claims.